DOCUMENT RESUME

ED 159 225 TM 007 709

AUTHOR	Bormuth, John R.
TITLE	Feasibility of Criterion-Referenced Testing and Evaluation.
PUB DATE	Jan 78
NOTE	42p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978); Not available in hard copy due to marginal legibility of original document
AVAILABLE FROM	John R. Bormuth, University of Chicago, 5835 S. Kimbark Avenue, Chicago, Illinois 60637 ($5.00 plus postage)
EDRS PRICE	MF-$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS	Achievement Tests; *Criterion Referenced Tests; *Cutting Scores; Elementary Secondary Education; Mastery Tests; Reading Tests; Statistical Analysis; Task Performance; *Test Bias; *Test Construction; *Testing Problems; Test Reliability; Test Validity; *Theories
IDENTIFIERS	*Test Theory

ABSTRACT

The feasibility of criterion referenced testing is held to be dependent on the tenability of two postulates: (1) that bias can be controlled in a principled manner from one test to the next; and (2) that one mental process measured by such tests may lawfully interact with another. Without the first postulate, criterion scores could not be established rationally. Without the second postulate, the usefulness of reaching the criterion score could not be established. These concepts apply both to content mastery tests and to task proficiency tests. Two corresponding research hypotheses are the hypothesis of regression identity for lawfully made tests and regression regularity. Results of earlier research on cloze tests, reading comprehension completion tests, and three preference scales relating to reading materials administered in grades 3-12 were discussed. High correlations were obtained between the cloze tests and the reading comprehension tests and curvilinear regressions between cloze tests and preference scales that confirmed the original hypotheses. These results were taken to confirm that cutting scores may be rationally determined and that algorithmically generated criterion referenced tests may be developed to measure significant educational outcomes. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

January, 1978.

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

FEASIBILITY OF CRITERION-REFERENCED TESTING AND EVALUATION

John R. Bormuth

University of Chicago

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY
John R. Bormuth
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

ED159225

Early in this century and before, many believed that education was supposed to serve the needs of society and that educators were accountable for demonstrating the effects they had on their students. In particular, as this belief went, society establishes schools to produce students who can perform certain real-world tasks competently; those tasks require the use of certain knowledges; those knowledges can be tested; and educators ought to demonstrate that a student has learned those knowledges before he is promoted from one grade to the next and that he can competently apply those knowledges to real-world tasks before he is granted a diploma implying that he can (Caldwell and Courtis, 1923).

Two kinds of tests were used for these purposes. One kind, call it a content-mastery test, was designed to inventory a student's knowledge of each item of instructional content that he was supposed to learn in a unit or course of instruction. And he was promoted from or retained in this instruction depending on whether he reached a criterion score that educators believed would forecast competent performance of real-world tasks. The other kind, call it a task-proficiency test, was designed to determine if a student could indeed apply his knowledge competently to real-world tasks, tasks such as reading the Constitution, translating a passage from Cicero, or operating a lathe. And many students were either failed or granted their diplomas, depending on whether their scores on these task-proficiency

TM007709

tests reached the criterion score deemed to represent competent performance.

One result of these practices was that school promotions and diplomas at least ostensibly meant what they implied, namely that students had mastered certain content and could apply their knowledge competently to certain real-world tasks. Unfortunately, another result was that lower grade classrooms were burdened and disrupted by a large number of over-aged and very frustrated students struggling to complete first, second, and third grade requirements (Billett, 1933). And that, in turn prompted many to rear up in humanitarian indignation to demand that schools recognize that students were different individuals and that the schools study and adapt their instruction to meet differences in aptitude, ambition, and temperament (Cook, 1941). Then the critics discredited the traditional practices by striking at their Achilles heel, at the testing and evaluation procedures on which they were based. They pointed out that educators had no firm evidence to support their beliefs that a given score on one of their tests always corresponded to a certain amount of knowledge or proficiency, that their content-mastery criterion scores did in fact forecast competent task-proficiency, and that their criterion score of task-proficiency did indeed represent the most desirable level of task performance (Tiegs, 1931, 1939; Starch, 1918). Educators' inability to respond left the strong presumption that test scores were arbitrarily confounded by the ideosyncracies of test writers, tasks, situations, and students, and that the criterion scores were arbitrary guesses. So the educator's sand castle was swept away before the tide of individual difference reformers.

The most basic reason why educators could not defend their practices was that they had no means for controlling test biases in a principled

manner from test to test. Consciously or unconsciously, an examiner could either pass or fail a student merely by choosing to test either obvious or obscure content and phrasing his questions in either simple or complicated ways. And no one knew how to formulate the rules needed to discipline sensibly his choices of questions. So, how could a criterion score forecast success on task-proficiency tests when it corresponded to no particular amount of knowledge in the first place? And how could a criterion score on a task-referenced test possibly be considered to be the most desirable one when it could correspond, willy-nilly, to just about any level of actual competence at performing those tasks. And, finally, by virtue of what natural law can we say that a given level of task-performance is regularly associated with quality of product, task enjoyment and the other things that might lead us to believe that a particular level of task performance is generally the most desirable one?

The movement to provide for individual differences arose after Binet's (Binet and Simon, 1905) introduction of his intelligence test and proceeded symbiotically with efforts of psychologists to develop the theory needed to measure and analyze the traits on which students differ. This body of theory views knowledge of content as a force-vector (Thurstone, 1935) and prescribes procedures for selecting items primarily with reference to their utility for representing this vector reliably and only secondarily and casually with reference to representing the domain of content items supposedly being tested. Then a student's score was evaluated not by comparing it to a criterion score but rather by comparing it to the distribution of scores made by his peers. This comparison supposedly allowed educators to adapt their instruction to fit student personal needs.

As most readers know, these vector-referenced, norm-evaluated tests and the pedagogical reasoning behind them still dominate instructional practices and have now brought us in full circle. For a new generation of critics has reared up in humanitarian indignation, pointed to the large number of illiterate and otherwise incompetent high school graduates and demanded that schools serve the needs of society, insist on mastery of content, and demonstrate their students' competence to perform real-world tasks. And somehow, the schools were supposed to do this without losing sight of individual differences in students.

In a seminal paper published in 1963, Robert Glaser suggested that the testing and evaluation needs of both causes could be served provided that credible theories could be developed to discipline the construction of content mastery and task-proficiency tests and to rationalize the selection of criterion scores. He named this body of theory criterion-referenced testing and evaluation. Some have rushed to restore the old content-mastery and task-proficiency practices back into practice, almost unchanged, ignoring the need for systematic research, ignoring the crushing arguments that can be leveled against those practices, and raising the odor of fad over the whole activity. Others, however, have set out to develop the systematic theory required (Hively et al, 1965; Bormuth, 1970; Finn, 1975) to control test bias and to identify rationally defensible criterion scores (Bormuth, 1971).

At the present time, the feasibility of this form of evaluation has come down to rest on two propositions, namely that, when tests are written

in a principled fashion, test biases are held essentially constant from one test to another and that the underlying knowledges and dispositions measured by such tests are lawfully associated with the consequences that determine the desirability of a test score. Some of the necessary test making procedures have been devised and put to use in ways that make it possible to put both propositions to preliminary tests. The next section explains the nature and relevance of these two propositions and uses them to deduce testable hypotheses. The third section puts those hypotheses to empirical tests. And the last section discusses some of the implications of their confirmation.

Rationale

Neither of these propositions can be tested directly since they refer partially to knowledges and other mental states that lie within a person's head. However, they can be tested indirectly. When they are taken together they can be made to imply still other propositions that are fully testable. Two of these implications furnished the research hypotheses examined in this paper. Call them the regression identity and regression regularity hypotheses. First we will consider the nature of each theoretical proposition and its relevance to criterion-referenced testing and evaluation. Then we will derive the two theorems that represent the two research hypotheses.

Theoretical Propositions

Classical achievement test theory recognizes that a test item is a

potentially biased indicator of the knowledge it purportedly tests. A person's response to an item is considered to be biased if he answers it incorrectly when in fact he has the knowledge tested or if he answers it correctly when in fact he does not have that knowledge. The item itself is biased when one type of these types of bias predominates in students' responses to it, and scores on a test are biased when its items are systematically biased.

Test scores are regarded as being directly comparable only when the major biasing agents are held constant in a test. Thus, for some examples, test scores are regarded as being comparable only if students of similar ages and backgrounds are tested under similar physical and psycho-social conditions using similar time limits, instructions, and items. In order for the items to be considered comparable, each set used must be formally identical to the others, meet a series of empirical statistical criteria, or be demonstrably similar in relevance to the instruction and in the manner of phrasing and selecting them. Demonstrating similarity of relevance, phrasing and selection lay largely beyond the means of classical test theory. But, since that theory was designed originally and primarily to compare students participating in the same instruction, bias could ordinarily be controlled by using the identical set of items for all or by using different sets that were matched statistically. And, although these tests were often used in efforts to compare students participating in materially different programs, strict classical theory eschews this practice on grounds that major biases are likely to arise from interactions among the contents of test items, the contents of instructional programs, and the traits of the students.

Criterion-referenced tests, however, are supposed to be used to compare not only students but real-world tasks and instruction that may differ substantially in content. This makes it implausible and often simply impossible to use the same set of items to test all the tasks being compared and has forced us to develop some fairly strict procedures for deriving comparable tests from materially different tasks. The hope is that by using the same rules to make tests from different tasks, and by making those rules sufficiently detailed, variances in scores from student-to-student and task-to-task will reflect only the different amounts of knowledge of the students and the different amounts of complexity of the tasks. Thus, scores on each test may be a biased indicator of the knowledge they measure, but this bias will be essentially the same for all students and tasks compared.

Principled Control of Test Bias: Thus, we arrive at the proposition that bias can be controlled in a principled manner from one test to the next. That is to say, if the relationships are held constant between the contents and forms of tasks and the contents and forms of the items in tests of those tasks, if the items included in the tests are sampled in the same way, and if the tests are administered to a group of students while holding constant all other major sources of test bias, then the amount of knowledge, x, that a student, s, obtains from or applies to a given task, t, is a function, f, of his score, X on that test, the function being essentially (i.e. allowing for the usual errors of measurement) the same for all students and tests in the set. That is to say,

$$x_{st} = f(X_{st}). \qquad (1)$$

And specifically, this function is not arbitrarily dependent on the particular student and task tested. Which is to say,

$$\sim[x_{st} = f_{st}(X_{st})],$$

wherein f would have to be subscripted to indicate its arbitrary dependence on the students and tasks that were tested.

This is not to deny, however, that processes that produce that test bias, like any other mental capabilities, change systematically with such things as the age, A, of the student or the complexity, C, of the task. Rather, it is to assert that those biasing processes, like any others, may be lawfully accounted for, person-to-person and test-to-test, under these testing conditions. Thus, it might very well be the case that

$$x_{st} = F(X_{st}, A_s, C_t, \ldots)$$

but that general functional relationship, F, is lawful across all students and materials tested in this manner.

Figure 1 may help to make clear the exact nature of the assertion that biases are controlled. Suppose that tests A and B were made from two dif-

---

Insert Figure 1 about here.

---

ferent pieces of instruction, while holding constant the algorithms by which the test items were selected, formatted, and phrased. Also, suppose that each test were given to different groups of students who were studying from those pieces of instruction, while holding constant the manner in which the tests were administered. And, finally, suppose that we had

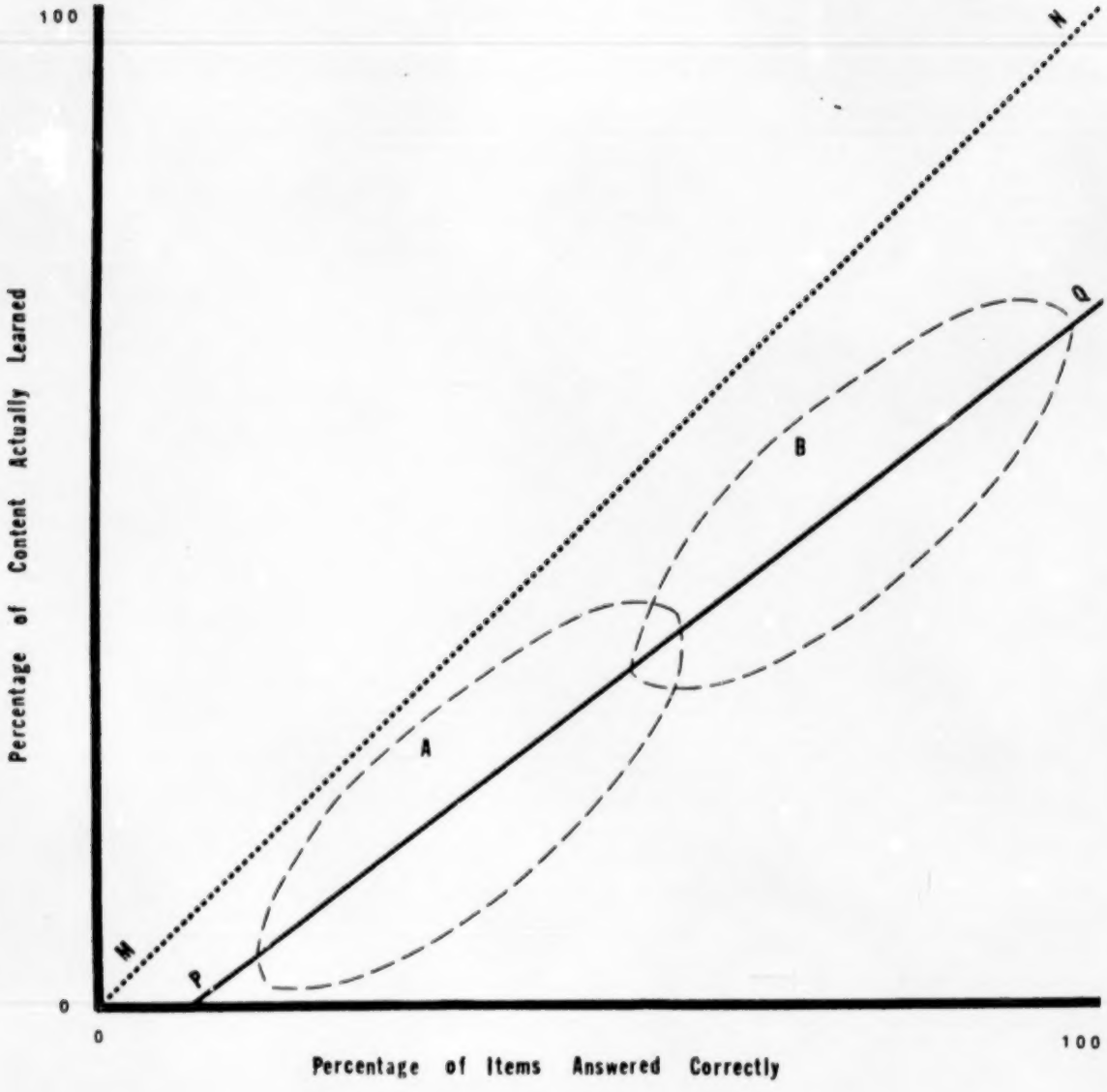

Figure 1. Illustration of the assumption of principled control of test bias.

somehow looked inside each student's head to determine exactly what content he had actually learned and then regressed his observed percentage score on the percentage of the content that he had actually learned. Now, common sense tells us that, if our controls had removed all biases in the two tests and if those tests contained no random error of measurement, then every pair of scores plotted would fall somewhere on line MN, which has an intercept of zero and a slope of one. That is, observed scores of 10 or 78 percent would in every case correspond to actually having acquired those identical percentages of the knowledges tested. But, in the real world, we would expect some uncontrolled sources of error to randomly disperse each set of scores about its regression line and some controlled sources of bias to cause the observed scores to shift the origins of those regression lines away from zero and their slopes away from unity, thereby causing the observed scores to systematically over- or underestimate the amount of content a student had actually learned.

Now, principled control of test bias assumes that all of the major biasing agents of this pair of tests were controlled in identical ways and that the sizes of their biases would, therefore, all have been identical. Therefore, it follows, as this postulate asserts, that the regression lines of both tests, and all other tests made and administered the same way, must lie along a single path, tracing out some single line, such as line PQ and varying from that line only within the limits one would expect, given the sizes of the random errors associated with sampling students and content items. All of which boils down to saying that when the biases in different tests are controlled in the same ways, then scores on those tests should

generally bear the same relationships to the underlying traits that they measure. Or, to put it in still a third way, scores on such tests should exhibit identical regressions on the traits that they measure.

Lawful Interaction of Processes: The second theoretical proposition is simply a generalized statement of the familiar proposition that one mental process may lawfully interact with another. Much psychological theory is related to multiple stage processes in which the nature and level of one process "determines" the nature and level of other processes. The Dember-Earl theory (Dember, 1965), for example, operationally relates indications of a subject's preferences for tasks to the complexity of those tasks, asserting (roughly speaking) that a subject prefers tasks that are of intermediate complexity for him to tasks that are either much more simple or complex. Then the theory explains this empirical fact by saying that the level of information obtained by cognitively processing the task serves as the effective stimulus for affective processes that, in turn, determine his level of liking for the task and hence his perseverence on, ratings of, or tendency to engage in the task.

That is to say, the level of a student's product, x, from performing one type of processing on a task may lawfully "determine' the level of product, y, that he obtains from performing another, more or less different, type of process on that task. Thus,

$$x_{st} = g(y_{st}),$$

which of course, as was also the case in principled control of bias, implies that

$$\sim[x_{st} = g_{st}(y_{st})],$$

but, also, recognizes the possibility that

$$x_{st} = G(y_{st}, A_s, C_t, \ldots)$$

Figure 2 may help to explain the nature of the assertion that processes interact lawfully. Suppose that we could somehow look into a student's head

---

Insert Figure 2 about here.

---

and determine the exact levels of learning and interest that he had experienced in performing a task. If these two processes interacted lawfully in the manner illustrated by the curve in this figure and if the task were fairly easy for him, placing him at point E, then he would also be most likely to exhibit a certain corresponding level of interest, placing him at point P on the curve. Then, if he were given successively harder tasks, his interest levels on them would be likely to rise and then fall along this curve, placing him at point Q when he reached level of difficulty H.

In actuality, of course, we would expect difficulty to only partially determine interest, causing a dispersion of interest levels at each level of difficulty. But the assumption that processes interact lawfully would require that the central trend of that relationship trace out some single path, such as this curve, for any student performing tasks of differing difficulties or for students of differing abilities taking the same task. Or, to put it another way, the assumption of lawful interaction asserts that

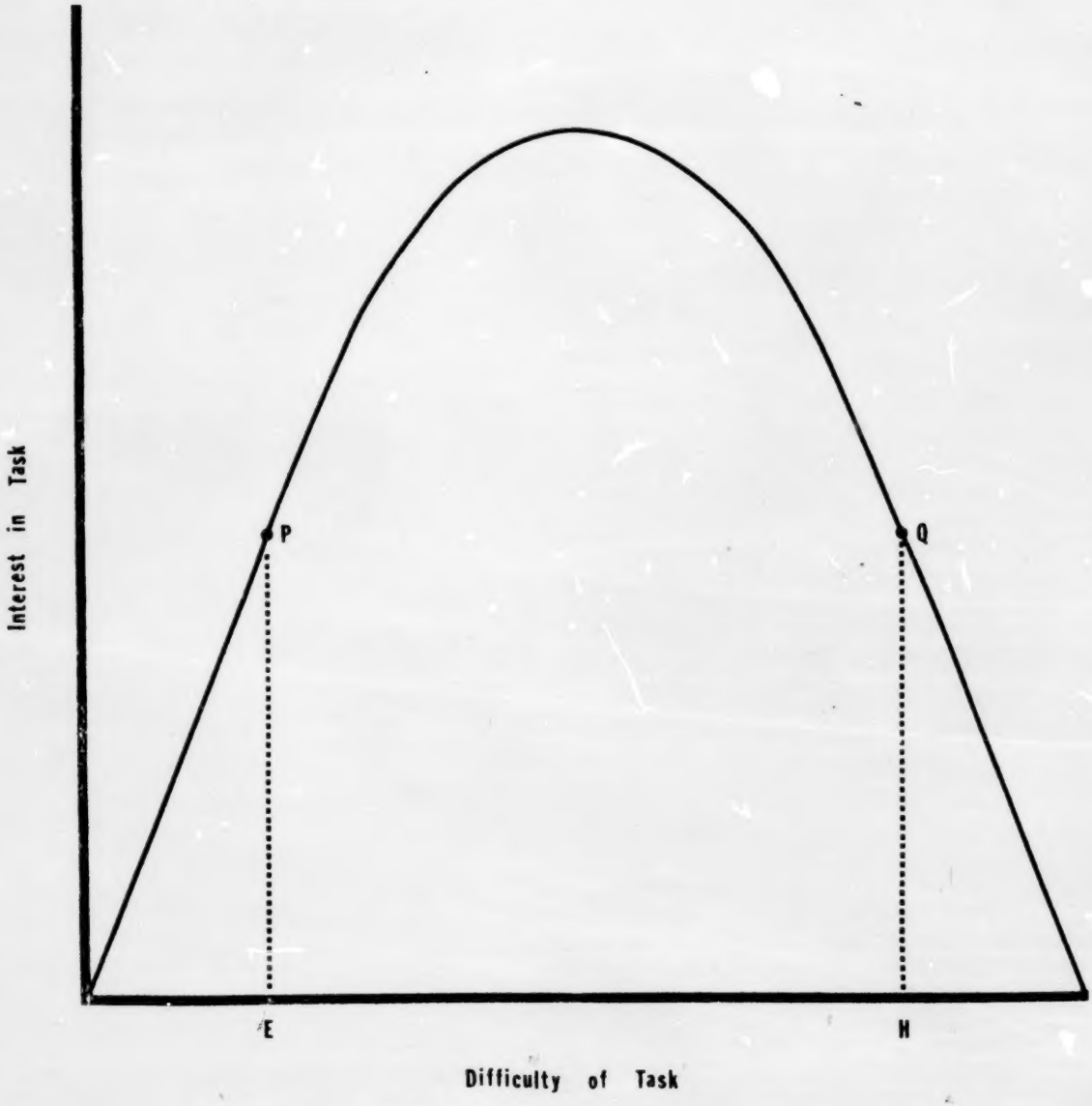

Figure 2. Illustration of the assumption of lawful interaction of process.

mental processes are related by laws of exactly the same sorts that Newton used to describe the relative motions of falling bodies and Boyle used to describe the behavior of gasses. Then, as we shall see next, when bias is controlled in a principled manner, the regressions between measures of interacting processes manifest and bear direct and constant relationships to those laws.

Research Hypotheses

Although neither of these two postulates is directly testable, they can be used to infer two hypotheses that are testable and whose truths can determine not only the truth of the two postulates but also the feasibility of criterion-referenced evaluation.

Regression Identity: The regression identity hypothesis asserts that, when pairs of different tests are made from each of several tasks, administered to the same students, and the scores on one test regressed on the scores of its pair-mate, those regression lines in the observed ranges of each set of scores will be essentially identical -- falling one on top of the other and tracing out a single path -- allowing of course for the usually expected amounts of error of prediction and stipulating that bias is controlled in a principled manner and the two processes interact lawfully. Thus, the proposition to be proven is that

$$X_{st} = i(Y_{st}), \qquad (3)$$

where X and Y are scores on tests of the products, x and y, of two more or less different processes and where it may be assumed that

$$x_{st} = f(X_{st}) \qquad (1)$$

$$y_{st} = h(Y_{st}) \qquad (1')$$

$$x_{st} = g(y_{st}). \qquad (2)$$

Substituting the right-hand members of (1) and (1') into (2) yields

$$f(X_{st}) = g(h(Y_{st})), \text{ or}$$

$$X_{st} = f^*(g(h(Y_{st}))).$$

Then, since f*, g and h are merely numeric parameters, they can be collected to form the simplified set, i. Therefore,

$$X_{st} = i(Y_{st}). \qquad (3)$$

Now, since both X and Y in (3) are fully observable test scores and i is a set of parameters that can be calculated from data, the regression identity hypothesis is fully testable. Furthermore, since regression identity seems to rest necessarily on the postulation of principled control of test bias and the lawful interaction of processes, then confirmation of regression identity, other conditions being met, indicates that both postulates are true and a disconfirmation indicates that either or both are false.

Regression Regularity: The regression regularity hypothesis asserts that, when a test of one kind is made from each of several tasks that differ in complexity, administered to the same group of students along with a single test of some other kind (such as a vector-referenced or content-mastery achievement test), and when that single test is regressed on each

of the task-proficiency tests, those regression lines will either be parallel to each other or converge in a lawful and interpretable manner, allowing of course for the usual amounts of error of prediction and stipulating that the bias is controlled in a principled manner in the task-proficiency tests and that the capabilities measured by the two kinds of tests interact lawfully. Thus, the proposition to be proven is that

$$Z_s = m(X_{st}, \bar{X}_t), \qquad (4)$$

wherein $\underline{Z}$ represents student's scores on the single test, $\underline{X}$ represents scores on task proficiency tests, and $\bar{\underline{X}}$ represents the means of the student's scores on the task-proficiency tests.

The new assumptions required in this case are variants of the principled control of bias and the lawful interaction assumptions, namely that

$$x_{st} = f(X_{st}) \qquad (1)$$

$$z_s = j(Z_s), \text{ and} \qquad (1')$$

$$z_s = k(x_{st}, \bar{x}_t). \qquad (2')$$

Assumption (1') represents the degenerate case of the principled control of bias assumption, referring as it does to only one test. Moreover, it is the standard assumption made for any test that the underlying trait being measured is a function, presumably monotonic and increasing, of the scores on that test. Assumption (2') represents a semi-degenerate case of the lawful interaction assumption. It asserts that, if $\underline{z}$ lawfully interacts with $\underline{x}$, then it will also interact lawfully with the mean levels of $\underline{x}$ that students would employ on differing tasks. Or, to put it another way, variations between amounts of $\underline{x}$ employed within a task represent variations in the same capability as variations between the mean amounts of $\underline{x}$

employed on tasks of differing difficulty. And finally, though the matter will not be developed here, it can be shown that this hypothesis amounts to asserting that the variances of the task-referenced tests either are equal or vary systematically with the means of those tests. Deriving (2') from (2) is also left as an exercise for the reader.

Now, substituting (1) and (1') into (2') and letting N represent the number of students who take these tests yields

$$j(Z_s) = k(f(X_{st}), \frac{1}{N}(\sum^{N} f(X_{st}))), \text{ or}$$

$$Z_s = j^*(k(f(X_{st}), \frac{1}{N}(\sum^{N} f(X_{st})))).$$

And then (assuming that f is linear) collecting j^*, k, and f into the single set of parameters, m, yields

$$Z_s = m(X_{st}, \bar{X}_t) \qquad (4)$$

Since the regression regularity hypothesis is a semidegenerate case of the regression identity hypothesis, an experimental test of its truth would not provide quite as strong a case for or against the assumptions of principled control of test bias and lawful interaction of traits. However, as will now be seen, it is equally crucial to the feasibility of criterion-referenced evaluation, and a disconfirmation of it would cast grave doubt on the viability of that type of evaluation.

Relevance of the Postulates

Both the postulates and the hypotheses play crucial roles in determining the feasibility of criterion-referenced testing and evaluation. The truth of the postulates is, of course, the decisive factor. But the hypotheses are also important for they refer to operations of the sort that

must be performed in order to identify rational criterion scores.

Postulates: First, consider the role of principled control of test bias: All distinctive uses of criterion-referenced testing make either explicit or implicit uses of a criterion score. The criterion score used ostensibly represents, all things considered, the most desirable level of performance on that test, a student's instruction being terminated if he exceeds it, for example, or his instruction being continued and revised if he fails to reach it. Now identifying a criterion score rationally is a very expensive operation that involves assessing the costs and benefits normally associated with each score on the test and assessing a decision maker's preferences for each of the factors comprising those costs and benefits (See Bormuth, 1971 , for example.) Hence, if bias cannot be controlled in a principled manner, answering a given percentage of the items on one test will correspond to a different level of actual mastery or proficiency than answering that percentage of items on other tests. And, no single score could be said to be the best for all tests of a given type; a different criterion score would have to be found for each, and the economic feasibility of criterion-referenced testing would have to be called into serious question.

Next, consider the role of lawful interaction of processes: In order to identify a criterion score, it is necessary to determine the important consequences that are normally associated with each score on the test. Some of the most important consequences associated with mastery and proficiency test scores are of a psychological nature. For example, a certain level of mastery might be preferred because it is regularly associated with a high level of transfer to and enjoyment of subsequent

instruction and thereupon lead to lower educational costs, perseverence in the course of study and likelihood of entering a profession that called for the use of the content studied. Now, if such processes do not interact lawfully, a score on one test may be associated with a very different consequence than the same score on even a test of the same type made from exactly the same instruction or task, requiring that a different criterion score be identified for every different test put into use and again calling into question the feasibility of criterion-referenced testing.

Hypotheses: Each of the hypotheses simulates an operation entailed in identifying criterion scores. First, the test of the regression identity hypothesis employed data that were actually used in an effort to identify criterion scores for interpreting one type of task-proficiency, comprehending didactic materials. This involved determining the amounts of information-gain, rate of reading, and student enjoyment of the reading task normally associated with a cloze score made from any such material. Much the same process would also be involved in identifying criterion scores for any task-proficiency or content-mastery tests. Hence, if regression identity does not hold, credible criterion scores cannot be identified. Next, the regression regularity hypothesis would normally come into play for identifying criterion scores for content-mastery tests. An important determinant of the value of a score on one of these tests is likely to be the proficiency it forecasts in dealing with real-world tasks of a sort that calls upon knowledge of the type of content being measured by this test. This can be determined relatively easily, but only if scores on the task-proficiency tests of this sort are related to content mastery scores in the simple manner described by the regression regularity hypothesis.

Evidence

The evidence bearing on these hypotheses came from two different studies, both of which are too complex to fully describe in this paper. So only the essential details will be given here and the reader referred to the originals (Bormuth 1971 and 1969) for details bearing on the regression identity hypothesis and the regression regularity hypothesis, respectively.

Regression Identity

The object of the study bearing on this hypothesis was to identify the most desirable score on cloze tests used to assess students' literacies with respect to verbal, didactic materials. Among several other things, this involved trying to determine how much information gain, rate of reading, and student preference was normally associated with each of the possible cloze percentage scores when the tests measuring these things were made from such materials. Although regression identity held for all the tests that were used in this study, several will be omitted here because of close similarity to one that is described or because the data on it was sparse.

Procedures: The tests used in this study were made from 32 passages, of from 250 to 300 words each, that were originally drawn in a larger sample (Bormuth 1969) to represent didactic materials used in 10 subject matter areas for the instruction of students in the primary grades through the college level. The materials represented eight levels of comprehensibility, ranging from the easiest to the most difficult found in the original sample. The four passages representing any one of these levels each dealt with a different discipline but were closely matched in comprehensibility as measured using cloze tests.

A cloze and a question-type comprehension test and four student pref-

erence scales were made from each of the passages. The cloze tests of 50 or more items each were made by replacing every fifth word with an underlined blank 15 spaces long and then scored counting a response as being correct only if it matched the word deleted. The question-type tests each consisted of 20 completion items made by applying verbation wh- transformations (Bormuth 1970) to randomly selected constituents in the sentence, anaphoric, and intersentence structures in a material and then scored, roughly speaking, by accepting responses that were either verbatim reproductions of the deleted constituent or an anaphorically indicated superordinate of it. Three preference scales asked the student to rate the material on a 1 through 7 scale for how well he liked the subject matter it dealt with, the style in which it was written, and his willingness to study it in instruction. A fourth scale asked the students to rate the suitability of the difficulty of a material for use in their instruction. It ranged from too easy to too hard. These ratings were folded around the midscore, 4, or the scale.

Ideally it was desirable to measure a student's performance on all tests made from a single passage, but this could not be done without incurring serious biases. To avoid those biases, the four passages within a difficulty level were rotated through four testing conditions in a latin squares design, giving a student the cloze test made from one passage, the question-type tests made from two passages, and the preference scales made from the remaining passage. The first question-type test was administered as a pre-reading or guessing test, not allowing the student to read the passage from which it was made, while the other question type test was administered post-reading in the usual way that comprehension tests are given. The scores on this set of tests were treated as if all came from

tests made from the same passage. None of the tests was timed.

The materials were then administered to 1600 students enrolled in grades 3 through 12 in a single school district that served a predominantly white, middle-income, suburban community. The design consisted of 8 levels of difficulty of materials with four passages nested within each level and 10 grade levels of students with 20 students nested within each difficulty and grade level.

Caution must be used in interpreting the sizes of reliability and correlation coefficients of task-referenced tests of the sort used here, and particularly those of the cognitive tests. The difficulties of their itmes were primarily determined by the complexity of the materials. Thus, the variances of some were sharply attenuated by ceiling and floor effects which in turn attenuated the reliability and correlation coefficients of the tests. For example, the reliabilities of the individual cloze tests ranged from .50 to .93 (pooling students across grade levels) and had the mean of .86 seen in the first column of Table 1. The reliability of a preference scale could be only crudely inferred by using the mean of its correlation with the same scale when students were asked to use it, first, to rate a material for use as a textbook and then again for use as a reference book. The mean of correlations between cloze and each of the other tests seen in the second column was obtained by averaging the correlations obtained within each difficulty level, but pooling students across grade levels.

Insert Table 1 about here.

Table 1. Reliabilities and correlations

Tests	Mean Values for Separate Tests[a]		Polynomial Multiple Regression[c]		
	Split-Half Reliability	Correlation[b] with Cloze	Cloze	Cloze, Grade	Cloze, Grade, Difficulty
Cloze	.86	---	---	---	---
Pre-reading	.61	.34	.74	.74	.80
Past-reading	.81	.64	.77	.78	.84
Subject Matter Preference	.71	.11	.11	.28	.30
Style Preference	.71	.14	.17	.33	.37
Difficulty Preference	.77	.21	.42	.42	.47
Willingness to Study	.74	.13	.11	.33	.34

a. Subjects pooled across grades and means calculated using Fisher Z transformation.

b. Calculated using significant ($p < .05$) terms from first three powers of cloze.

c. Calculated using significant ($p < .05$ terms selected from first eight powers of cloze scores, first three powers of grade and difficulty levels, and all interactions among terms involving different variables.

Analysis: First, consider the statistical evidence. Obviously, the first prerequisites that must be met in order for the regression identity hypothesis to hold between two types of tests is that those tests have some reliable variance and that, taken passage by passage, they have correlations of the same sign. That this was generally the case may be seen from the first two columns of Table 1. However, meeting just these criteria does not assure that regression identity holds, for it remains possible that the means and variances of the individual tests could have varied in ways that assured that their regressions followed anything but an identical path.

A stronger test of the hypothesis could be obtained by first pooling the data not only across grade levels but also across passage difficulty levels and then calculating a single correlation between the cloze tests and one of the other tests, treating all scores as if they came from a single pair of tests made from the same passage. If that correlation were not significant, the truth of the regression identity hypothesis would then be rendered problematic.

But, this third set of correlations could be either higher or lower than the set of average correlations seen in the second column of the table, depending on whether the error variances arising from the sampling differences associated with drawing different groups of students and passages were smaller or larger than the effects of increasing the variances of the cloze, and possibly the other, test scores due to the pooling of scores from tests made from passages of differing difficulty levels.

The size of this latter effect would, of course, depend on how similar the cloze test was to the other type of test. If, as was the case with the

pre-reading and post-reading tests, they measured fairly similar processes, varying the cloze difficulty of the passages by pooling them would substantially increase the variances of both sets of scores and probably increase the sizes of their correlations. And further, adding student grade level and passage difficulty to the regression model would probably increase the correlation only slightly since the information that those variables would carry about the probable level of a student's score on both tests would already be largely accounted for by his scores on this pair of tests. A comparison of the second column with the third and remaining columns in Table 1 shows that this was exactly what did happen to the correlations of the cloze scores with the scores on the pre-reading and post-reading tests. The correlation with the cloze scores rose substantially and adding grade and difficulty levels to the model had relatively smaller effects that were largely confined to the interaction terms in the regression model. Hence, regression identity seemingly held among these three kinds of tests.

Pooling across difficulty levels should have had less effect on the correlations between the cloze scores and the preference ratings since their correlations within difficulty levels were smaller to begin with. Pooling would increase variance of the cloze scores but would have little effect on the preference ratings. Hence any increase would be slight and roughly proportional to the size of the correlations within difficulty levels. The developmental effects represented by grade level could exert a stronger effect when added to the model, but difficulty level would, again, have little effect owing to its redundancy with the information contained in the cloze scores on individual tests. The correlations in the last four columns of Table 1 confirm these expectations and indicate that regression identity held among these tests as well.

Probably the most persuasive evidence, however, is obtained by inspecting plots of the regression curves showing performances on the question tests and the preference scales as functions of the cloze scores and grade levels of the students for the pooled data.

Insert Figures 3-8 about here.

Only the salient features of these curves will be mentioned here since their forms have been interpreted in detail elsewhere (Bormuth, 1971) and seen to agree with predictions based on the assumption of regression identity and existing psychological theory. The scores on the question tests, Figures 3 and 4, rose much as one would normally observe when pairs of these tests are made from a single test. The grade level effect resulted chiefly from increases in the variances of the question type tests, increases that could have been due to students' development of testwiseness and mathemagenic processes. The nature of the theory relating to the difficulty preference scales, Figure 7, is too complex to indicate here, but the centrally relevant fact is that these curves were orderly. Finally, the curves for the remaining preference scales, Figures 6 and 8, conformed reasonably well to the predictions of the Dember-Earl theory mentioned earlier. And the decrease in ratings associated with grade level could be attributed to the same factors that cause older students to become increasingly blase about instruction.

Regression Regularity

The object of the study bearing on the regression regularity hypothesis was to develop some readability formulas. The study entailed selecting 330 passages, of about 110 words each, in equal numbers from didactic materials

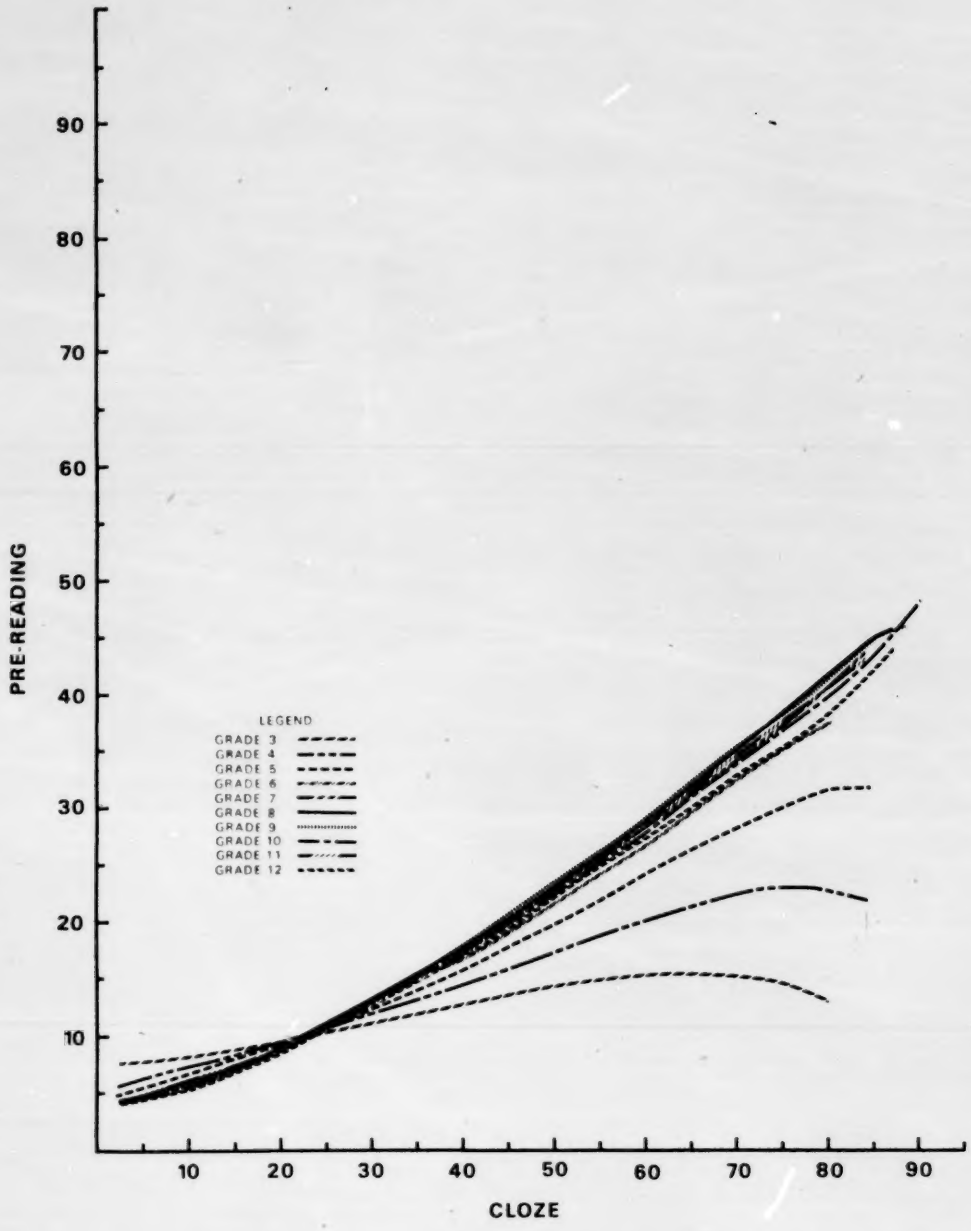

Figure 3. Pre-reading or guessing scores as a function of cloze scores and grade levels, pooling across levels of passage difficulty.

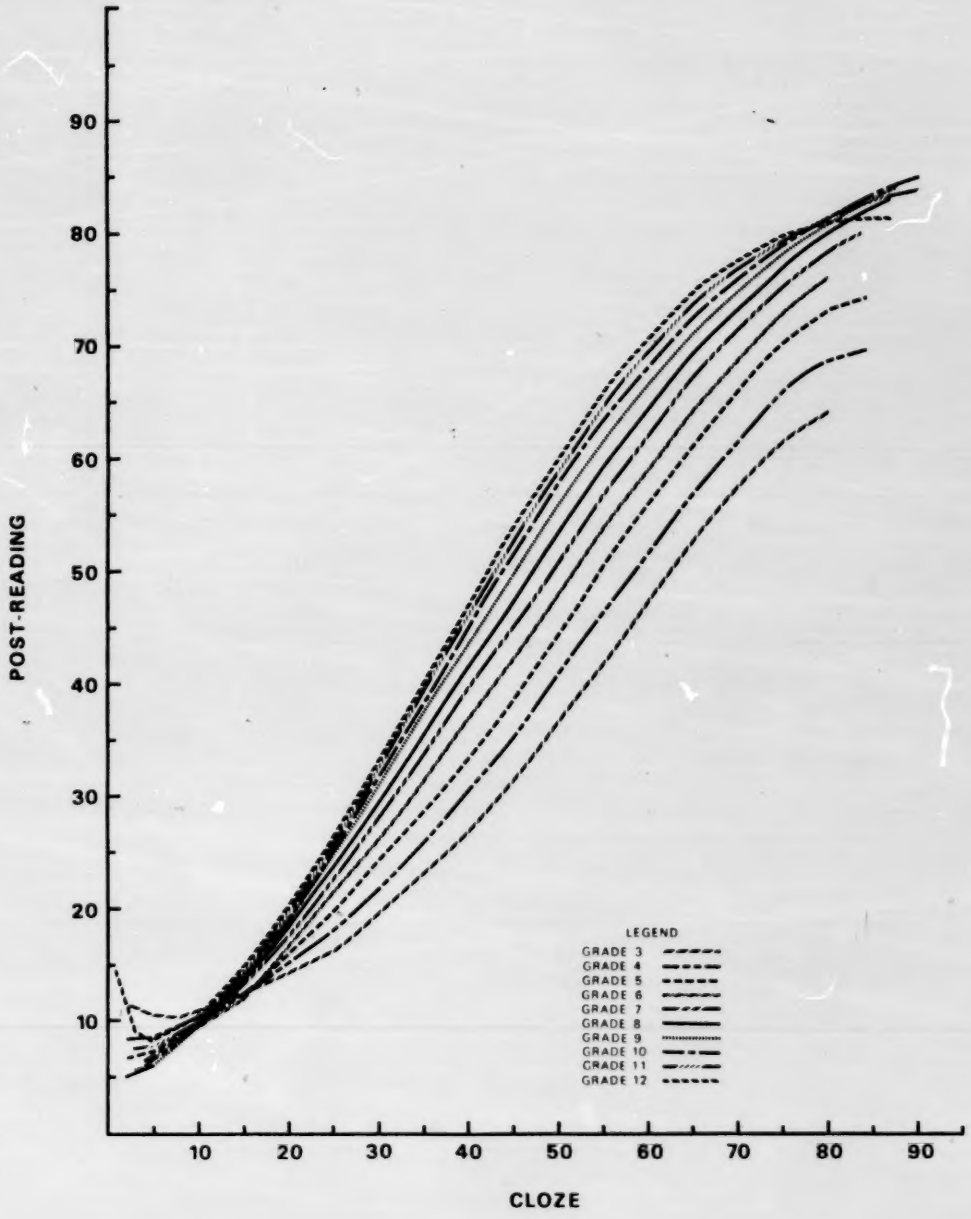

Figure 4. Post-reading or comprehension scores as a function of cloze scores and grade levels, pooling across levels of passage difficulty.

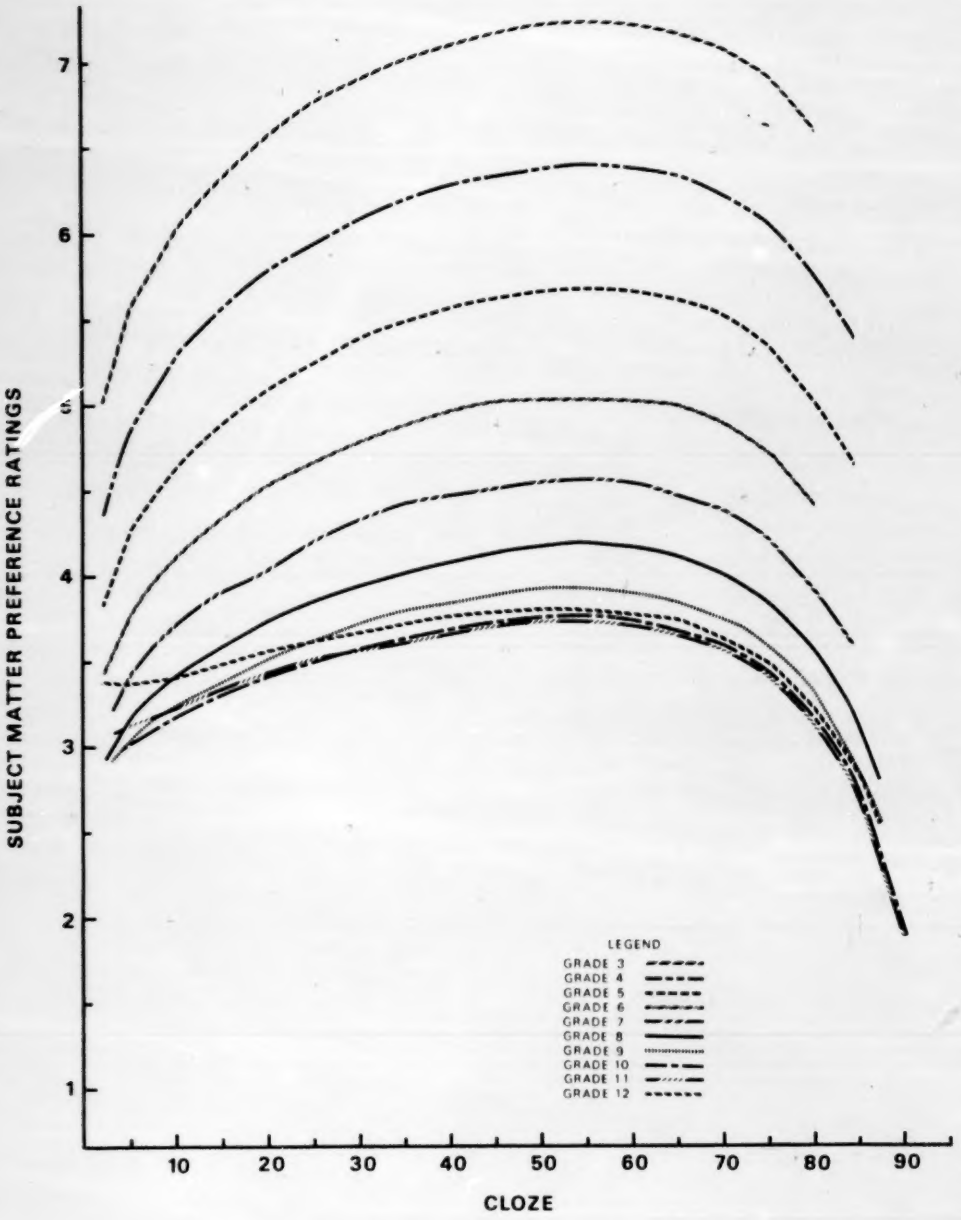

Figure 5. Subject matter preference ratings as a function of cloze scores and grade levels, pooling across levels of passage difficulty.

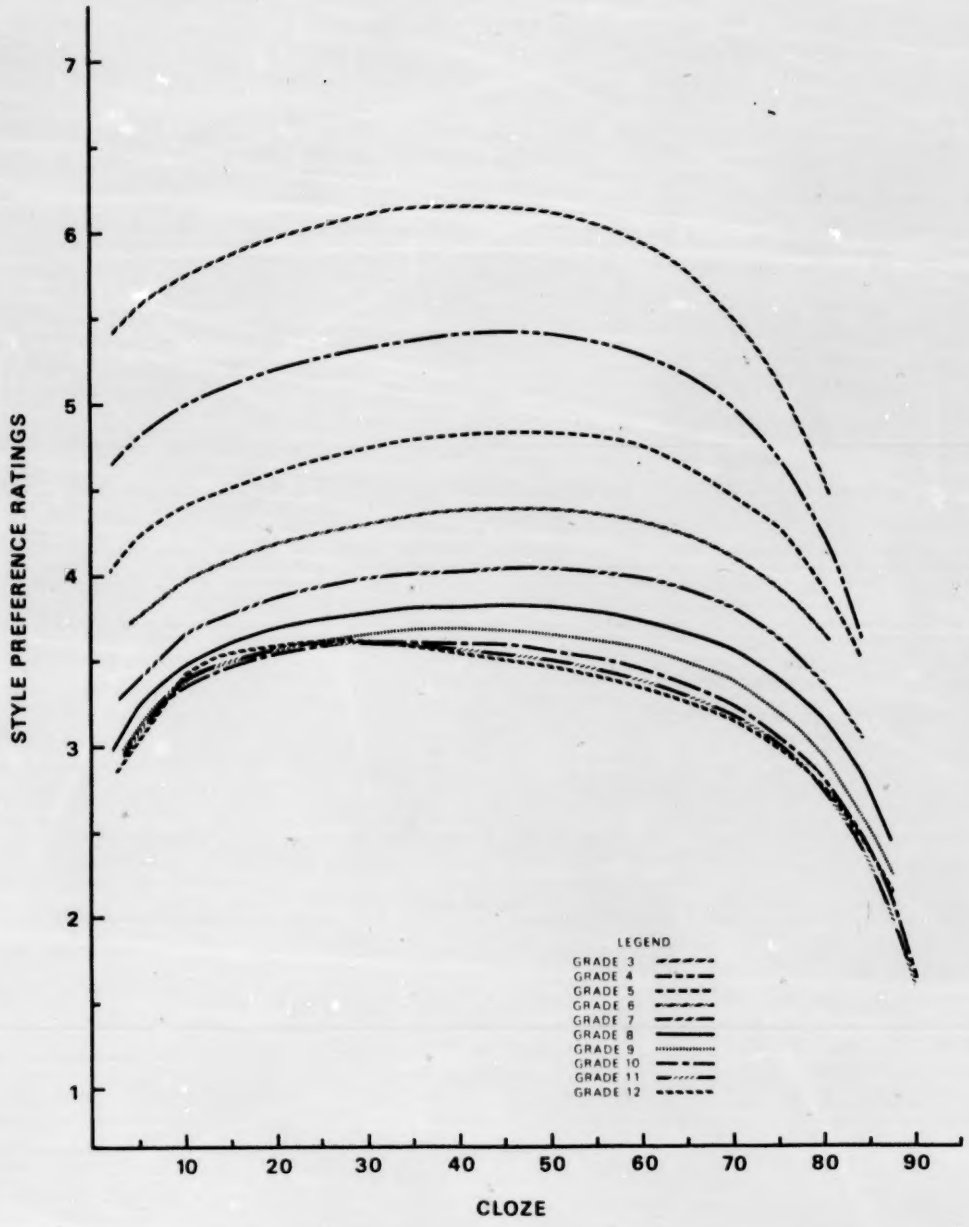

Figure 6. Style preference ratings as a function of cloze scores and grade levels, pooling across levels of passage difficulty.

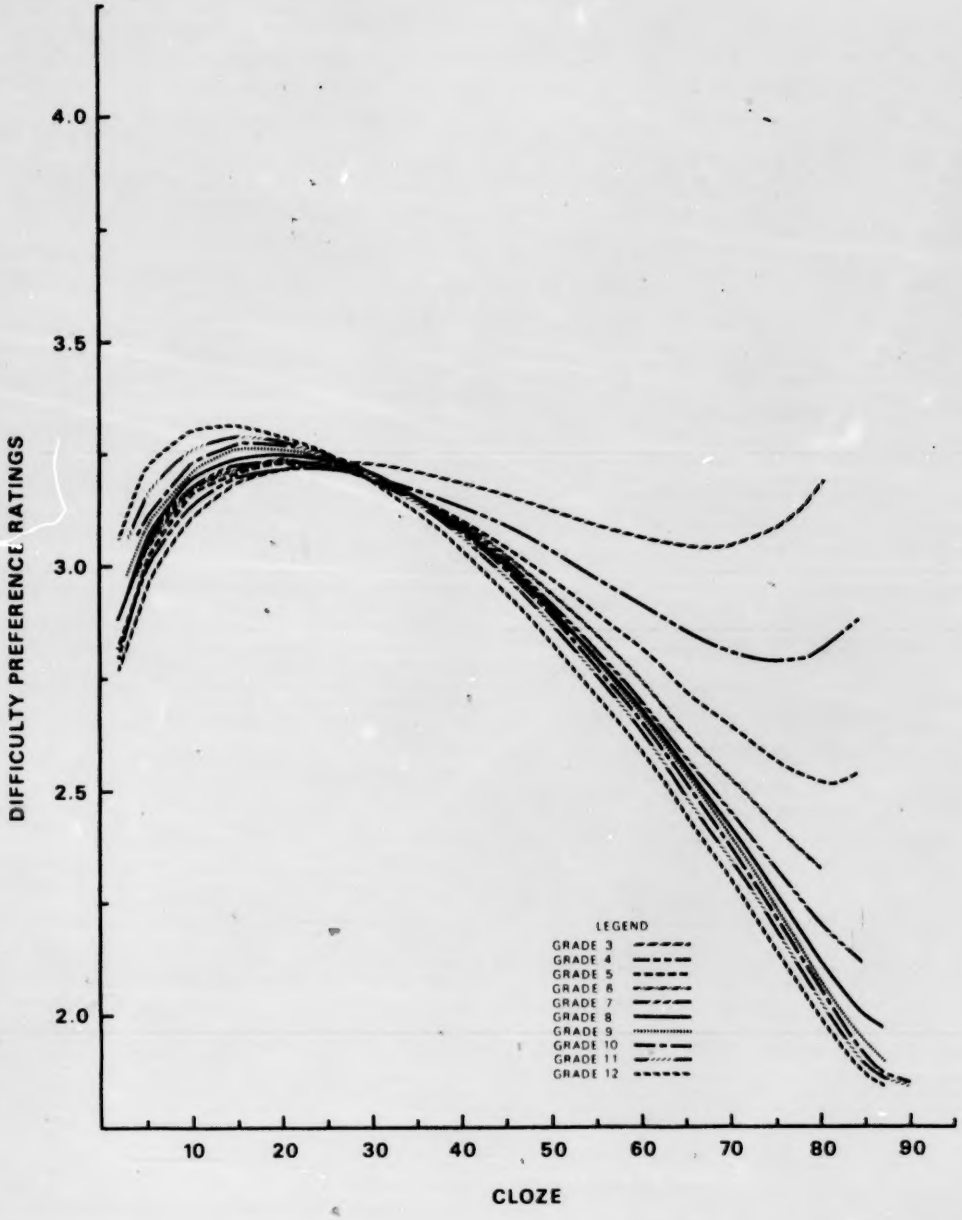

Figure 7. Difficulty preference ratings as a function of cloze scores and grade levels, pooling across levels of passage difficulty.

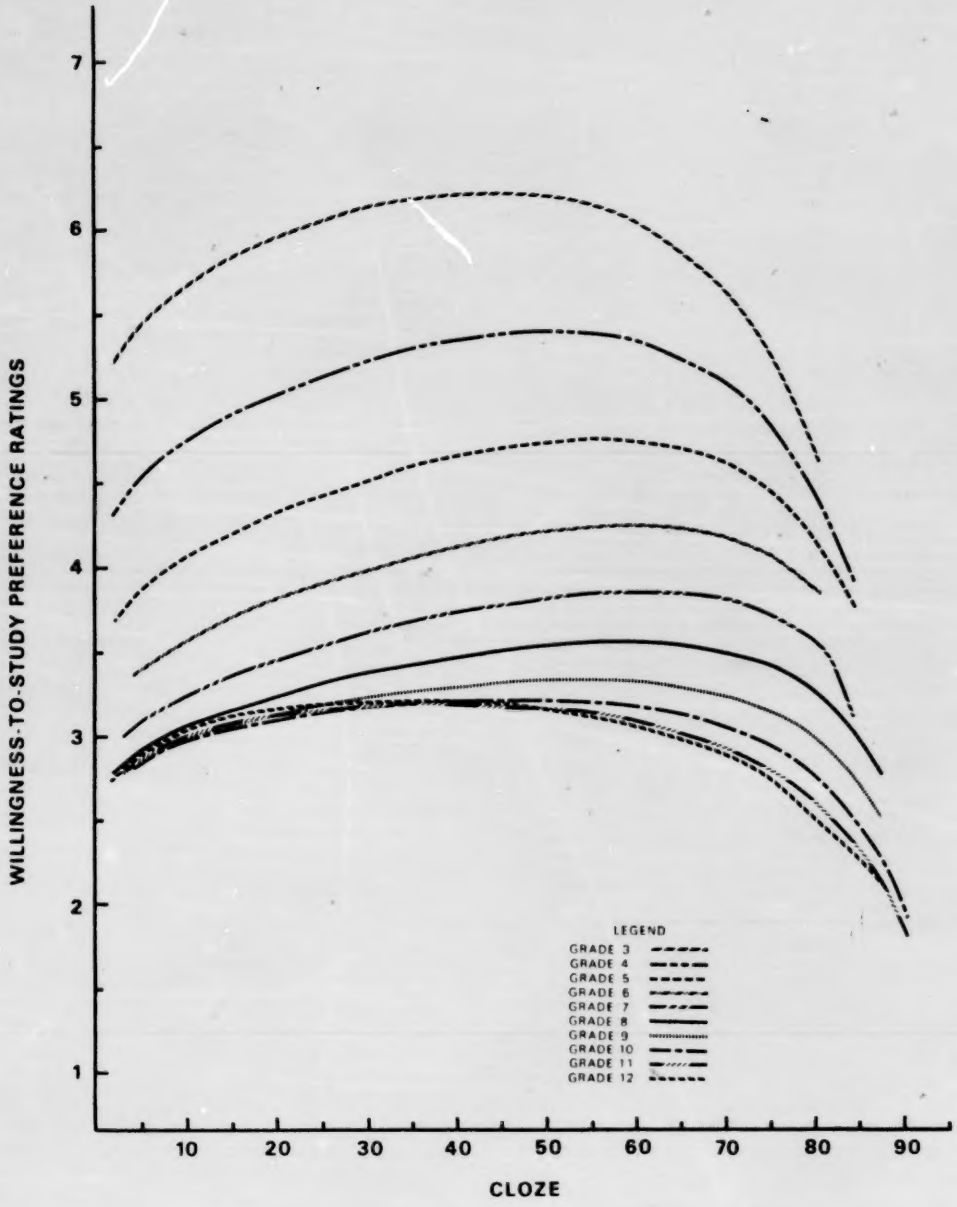

Figure 8. Willingness to study ratings plotted as a function of cloze scores and levels of difficulty, pooling across levels of difficulty.

in ten subject matter areas, stratifying on levels of school usage which ranged from primary grade through advanced college level. Five forms of a cloze test were made from each and then the California Reading Test and the cloze tests were administered to students ranging from grades 4 through 12, distributing the cloze tests and cloze test forms in equal numbers among 50 groups, with roughly 165 students in each, matched for grade level and reading achievement using a randomized blocking procedure. Since the cloze tests were fairly short, averaging about 21 items each, their reliabilities were fairly low, averaging about .65.

According to the regression regularity hypothesis, a student's grade equivalent score on the California Reading Test should be almost entirely predictable from his score on one of the cloze tests and the mean of the cloze scores on that passage. In addition, the results of the foregoing study suggested that, as students advance through the grades, they develop skills that are especially adapted to dealing with question type tests and have little to do with comprehending actual materials as that is measured using cloze tests. Since commercial achievement tests are composed largely of questions, it seemed advisable to remove grade level as a rival hypothesis by conditioning this regression on grade level. Hence, the model fit to the data attempted to predict a student's California Reading Test, grade-equivalent score, A, as a function of his grade level, G, his score on a cloze test, C, the mean of the students' cloze scores on that passage, $\bar{C}$, and all of the possible interactions among these terms.

Only the main effects for grade level, cloze score, and cloze mean and the cloze score by cloze mean interaction accounted for significant ($p < .01$) amounts of the variance. The regression equation, then, was

$$A = 2.24 + .832G + 7.00C - 3.79\bar{C} - 4.02\ C\bar{C}. \qquad (1)$$

The multiple correlation coefficient was .89 and would undoubtedly have been considerably higher had the cloze tests been longer and more reliable. Hence, this regression analysis strongly affirmed the truth of the regression regularity hypothesis.

Insert Figure 9 about here.

Figure 9 shows a plot of this regression equation for students in grades 4 and 12. The main effect for grade level simply shifted the regressions of the cloze tests upward, suggesting that a student could achieve increasingly high scores on the reading achievement test simply by advancing from grade to grade and without gaining in his ability to comprehend actual materials as that is measured by cloze tests. The main effects of cloze scores and cloze means were to "increase" the student's reading achievement scores, and the effects of the cloze score by cloze mean interaction was to increase the slope of the regression of cloze scores on the achievement test scores as the cloze means fell. Or to describe this interaction another way, the variances of the cloze scores decreased along with the means of the tests on which those scores were observed.

Discussion

These data, as far as they went, affirmed the truths of the regression identity and regularity hypotheses and, by implication, the two theoretical propositions that principled control of test bias could be achieved and that the processes measured by different tests could interact lawfully. The questions that remain to be answered are how far these results can be gen-

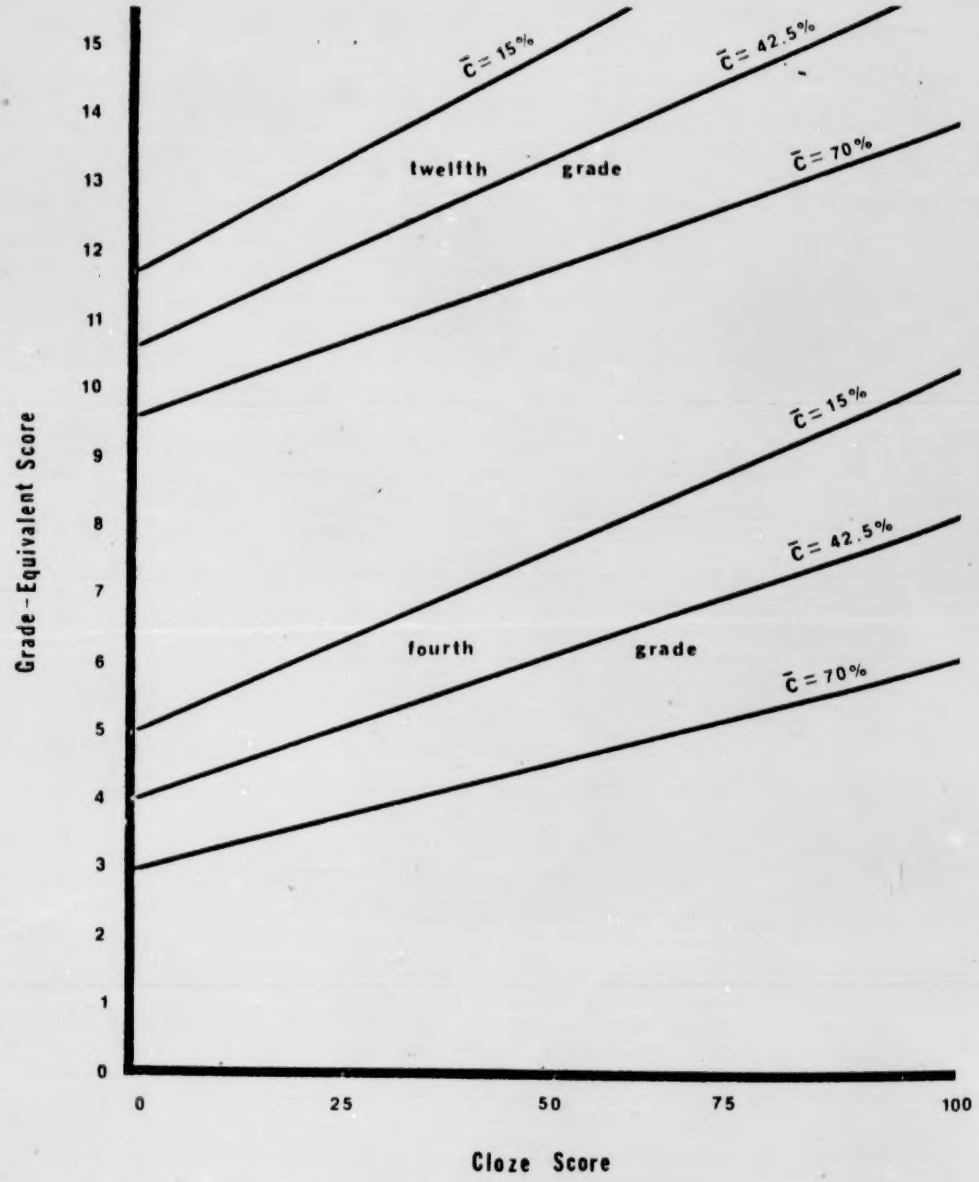

Figure 9. Reading Achievement test scores as a function of cloze scores, cloze test means, and grade levels.

eralized and what might be their significance for criterion-referenced testing and evaluation, for psychological research methodology, and for research in testing methodology.

Strictly speaking, these results can be generalized only to task-proficiency tests composed of items that have been generated from verbal, printed materials and selected from those generated using well defined algorithms. Further, these results make at least the strong suggestion that the same things would hold true for similarly derived content-mastery tests prepared and used for evaluating specific, well-defined instructional treatments of a verbal nature. But they do not generalize at all to casually defined, objective-referenced tests or to any other tests that purport to yield comparable measures for students who may have undergone different instructional treatments dealing in a variety of ways with the content supposedly represented by the tests. Indeed, in view of the large interactions possible between the form and factual content of such tests and the form and factual contents of the varieties of instruction, there are strong a priori grounds for doubting that either of the hypotheses or the theoretical postulates on which they rest could hold true for such tests. Fortunately, the proposal for conducting criterion referenced evaluation does not necessitate the use of casually made tests.

Next, consider the implications for evaluation. Earlier in this paper, it was claimed that one could not rationally evaluate instruction without measuring students' proficiency at performing real-world tasks, that doing so necessitated obtaining proficiency measures that were comparable from task to task, and that obtaining such measures required that we achieve principled control of test bias. Then it was argued that establishing

rationally plausible criterion scores on either task-proficiency or content mastery tests necessitated showing that each score on tests of these kinds was lawfully associated with certain amounts of each of the consequences, including other mental processes, that would cause people to place more value on one score than on another, and that for these associations to be lawful requires not only the principled control of test bias but also the lawful interaction of processes. Since the present study went some way toward confirming the truths of these two theoretical propositions, it also went some way toward confirming the feasibility of criterion referenced testing.

Third, these confirmations of the principled control of test bias and the lawful interaction of processes suggest that it is possible now to develop empirically disciplined theories about how processes interact as people perform laboratory and real world tasks. As mentioned earlier, conventional wisdom holds that cognitive and affective processes interact lawfully among themselves and with each other during the performance of tasks. But in the absence of a measurement methodology for relating the level of one process with that of another, these beliefs have been tested only qualitatively, if at all. It now seems evident that carefully made task referenced tests can provide a "window" through which to examine those interactions in detail. The view provided by any one test and group of students may, by itself, be a narrow one. Thus, for example, fourth grade students taking a cloze and question test on a hard passage would reveal only the southwest end of the curve shown in Figure 4. But varying the difficulty of the passage and the ability of the students, barring aptitude or age effects, would reveal the full form of the law governing the interaction of the

processes measured by those tests. Further, since it is possible to control test bias, it is at least conceivable that it could be studied and removed, making it possible not only to determine the general form of these laws but also to calculate their exact parameters.

Fourth, there are two highly suggestive and potentially important parallels between the methodological theory that led to algorithmically generated tests and the measurement theory that has always underlain test designs in experimental psychology. For one parallel, both insist that inferences about the nature of the processes underlying a response to a test task are moot unless one can correlate the content and structure of the test task and response with the contentent and structure of the original instructional stimulus. For one example, whether the response DAX to the stimulus BOF represents a rote or a transfer process depends on the relationship of DAX and BOF to the corresponding item in the instruction list. In parallel, what process is inferred when green is given in response to What color is Joe's house? depends on whether the subject was told Joe's house is green or Joe's house blends with summer foliage. For the other parallel, applications of both methodologies appear to rest always on the assumption of principled control of test bias and sometimes on the assumption of lawful interaction of processes. Exploring these two parallels in depth might lead to a better understanding of the basic logic underlying the inference systems of both branches of psychology and to the synthesis that would close the much lamented chasm (Cronbach, 1957) that has long divided these two branches of psychological theory.

Finally, several (Hively, et al. 1973, Bormuth, 1970, and Finn 1975) have proposed that principled control of test bias could be achieved and

have developed and demonstrated the use of item generation and selection algorithms for doing so. Even learning to use one of these algorithms can require a large investiment in training in linguistics, logic, mathematics, or the like, and developing one of these algorithms requires even larger and more varied investments. So people have been understandably reluctant to make these investments until there was substantial evidence that such tests do control test bias and that there was much to be gained by doing so. Hopefully these studies will have gone some way toward arguing the case that doing so is worth while and toward furnishing the evidence that it can be done.

References

Billett, R.O., Provision for Individual Differences, Marking and Promotion. National Survey of Secondary Education, Monograph No. 13, Washington, D.C.: U.S. Office of Education, 1933.

Bormuth, J.R., Development of Readability Analyses. (USOE Tech. Rep. No. 7-0052), Chicago: University of Chicago, 1969.

Bormuth, J.R. On the Theory of Achievement Test Items. Chicago: University of Chicago Press, 1970.

Bormuth, J.R., Development of Standards of Readability: Toward a Rational Criterion of Passage Performance. Report of U.S. Office of Education Project No. 9-0237, University of Chicago, 1971.

Caldwell, O.W. and Courtis, S.A., Then and Now in Education. Yonkers-on-Hudson: World Book, 1923.

Cook, W.W., Grouping and Promotion. Series on Individualization of Instruction, No. 2, Minneapolis: University of Minnesota Press, 1941.

Cronbach, L.J., The two disciplines of scientific psychology. American Psychologist, 1957, 12, 671-681.

Dember, W.N., A new look at motivation. American Scientist, 1965, 53, 409-427.

Finn, P.J., A question writing algorithm. Journal of Reading Behavior, 1975, 7, 341-367.

Glaser, R. Instructional technology and the measurement of learning: Some questions. American Psychologist, 1963, 18, 519-521.

Hively, W. III, Patterson, H.L. and Page, S.H. Generalizability of performance by Job Corps trainees on a universe-defined system of achievement tests in elementary mathematical calculation. Paper read at the American Educational Research Association Convention, February, 1965.

Starch, D., Educational Measurement. New York: Macmillan, 1918.

Thurstone, L.L., The Vectors of the Mind. Chicago: University of Chicago Press, 1935.

Tiegs, E.W., Tests and Measurements in the Improvement of Learning. Boston: Houghton Mifflin, 1939.

Tiegs, E.W., Tests and Measurements for Teachers. Boston: Houghton Mifflin, 1931.